The Reionization of Cosmic Hydrogen by the First Galaxies

Abraham Loeb
Department of Astronomy, Harvard University, 60 Garden St., Cambridge MA, 02138

Abstract

Cosmology is by now a mature experimental science. We are privileged to live at a time when the story of genesis (how the Universe started and developed) can be critically explored by direct observations. Looking deep into the Universe through powerful telescopes, we can see images of the Universe when it was younger because of the finite time it takes light to travel to us from distant sources.

Existing data sets include an image of the Universe when it was 0.4 million years old (in the form of the cosmic microwave background), as well as images of individual galaxies when the Universe was older than a billion years. But there is a serious challenge: in between these two epochs was a period when the Universe was dark, stars had not yet formed, and the cosmic microwave background no longer traced the distribution of matter. And this is precisely the most interesting period, when the primordial soup evolved into the rich zoo of objects we now see.

The observers are moving ahead along several fronts. The first involves the construction of large infrared telescopes on the ground and in space, that will provide us with new photos of the first galaxies. Current plans include ground-based telescopes which are 24-42 meter in diameter, and NASA's successor to the Hubble Space Telescope, called the James Webb Space Telescope. In addition, several observational groups around the globe are constructing radio arrays that will be capable of mapping the three-dimensional distribution of cosmic hydrogen in the infant Universe. These arrays are aiming to detect the long-wavelength (redshifted 21-cm) radio emission from hydrogen atoms. The images from these antenna arrays will reveal how the non-uniform distribution of neutral hydrogen evolved with cosmic time and eventually was extinguished by the ultra-violet radiation from the first galaxies. Theoretical research has focused in recent years on predicting the expected signals for the above instruments and motivating these ambitious observational projects.

1 Introduction

1.1 Observing our past

When we look at our image reflected off a mirror at a distance of 1 meter, we see the way we looked 6.7 nanoseconds ago, the light travel time to the mirror and back. If the mirror is spaced 10^{19} cm $\simeq$ 3 pc away, we will see the way we looked twenty one years ago. Light propagates at a finite speed, and so by observing distant regions, we are able to see what the Universe looked like in the past, a light travel time ago (Figure 1). The statistical homogeneity of the Universe on large scales guarantees that what we see far away is a fair statistical representation of the conditions that were present in in our region of the Universe a long time ago.

This fortunate situation makes cosmology an empirical science. We do not need to guess how the Universe evolved. Using telescopes we can simply see how it appeared at earlier cosmic times. In principle, this allows the entire 13.7 billion year cosmic history of our universe to be reconstructed by surveying the galaxies and other sources of light to large distances (Figure 2). Since a greater distance means a fainter flux from a source of a fixed luminosity, the observation of the earliest sources of light requires the development of sensitive instruments and poses challenges to observers.

As the universe expands, photon wavelengths get stretched as well. The factor by which the observed wavelength is increased (i.e. shifted towards the red) relative to the emitted one is denoted by $(1+z)$, where z is the cosmological redshift. Astronomers use the known emission patterns of hydrogen and other chemical elements in the spectrum of each galaxy to measure z. This then implies that the universe has expanded by a factor of $(1+z)$ in linear dimension since the galaxy emitted the observed light, and cosmologists can calculate the corresponding distance and cosmic age for the source galaxy. Large telescopes have allowed astronomers to observe faint galaxies that are so far away that we see them more than twelve billion years back in time. Thus, we know directly that galaxies were in existence as early as 500 million years after the Big Bang, at a redshift of $z \sim 10$ or higher.

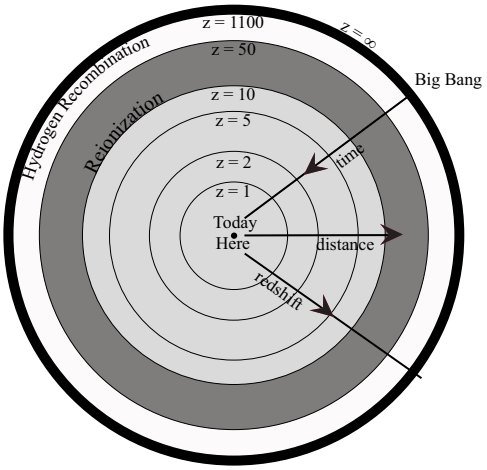

Figure 1: Cosmic archaeology of the observable volume of the Universe, in comoving coordinates (which factor out the cosmic expansion). The outermost observable boundary ($z = \infty$) marks the comoving distance that light has travelled since the Big Bang. Future observatories aim to map most of the observable volume of our Universe, and improve dramatically the statistical information we have about the density fluctuations within it. Existing data on the CMB probes mainly a very thin shell at the hydrogen recombination epoch ($z \sim 10^3$, beyond which the Universe is opaque), and current large-scale galaxy surveys map only a small region near us at the center of the diagram. The formation epoch of the first galaxies that culminated with hydrogen reionization at a redshift $z \sim 10$ is shaded grey. Note that the comoving volume out to any of these redshifts scales as the distance cubed. **Figure credit:** Loeb, A., *"How Did the First Stars and Galaxies Form?"*, Princeton University Press (2010).

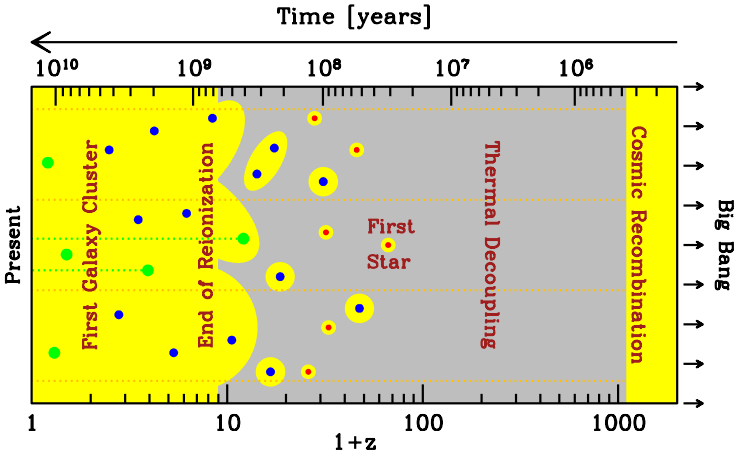

Figure 2: Overview of cosmic history, with the age of the universe shown on the top axis and the corresponding redshift on the bottom axis. Yellow represents regions where the hydrogen is ionized, and gray, neutral regions. Stars form in galaxies located within dark matter concentrations whose typical mass grows with time, starting with $\sim 10^5 M_\odot$ (red circles) for the host of the first star, rising to 10^7–$10^9 M_\odot$ (blue circles) for the sources of reionization, and reaching $\sim 10^{12} M_\odot$ (green circles) for present-day galaxies like our own Milky Way. Astronomers probe the evolution of the cosmic gas using the absorption of background light (dotted lines) by atomic hydrogen along the line of sight. The classical technique uses absorption by the Lyman-α resonance of hydrogen of the light from bright quasars located within massive galaxies, while a new type of astronomical observation will use the 21-cm line of hydrogen with the cosmic microwave background as the background source. **Figure credit:** Barkana, R., & Loeb, A., *Rep. Prog. Phys.* **70**, 627 (2007).

We can in principle image the Universe only if it is transparent. Earlier than 400 000 years after the big bang, the cosmic hydrogen was broken into its constituent electrons and protons (i.e. "ionized") and the Universe was opaque to scattering by the free electrons in the dense plasma. Thus, telescopes cannot be used to electromagnetically image the infant Universe at earlier times (or redshifts $> 10^3$). The earliest possible image of the Universe was recorded by the COBE and WMAP satellites, which measured the temperature distribution of the cosmic microwave background (CMB) on the sky.

The CMB, the relic radiation from the hot, dense beginning of the universe, is indeed another major probe of observational cosmology. The universe cools as it expands, so it was initially far denser and hotter than it is today. For hundreds of thousands of years the cosmic gas consisted of a plasma of free protons and electrons, and a slight mix of light nuclei, sustained by the intense thermal motion of these particles. Just like the plasma in our own Sun, the ancient cosmic plasma emitted and scattered a strong field of visible and ultraviolet photons. As mentioned above, about 400 000 years after the Big Bang the temperature of the universe dipped for the first time below a few thousand degrees Kelvin. The protons and electrons were now moving slowly enough that they could attract each other and form hydrogen atoms, in a process known as cosmic recombination. With the scattering of the energetic photons now much reduced, the photons continued traveling in straight lines, mostly undisturbed except that cosmic expansion has redshifted their wavelength into the microwave regime today. The emission temperature of the observed spectrum of these CMB photons is the same in all directions to one part in 100 000, which reveals that conditions were nearly uniform in the early universe.

It was just before the moment of cosmic recombination (when matter started to dominate in energy density over radiation) that gravity started to amplify the tiny fluctuations in temperature and density observed in the CMB data. Regions that started out slightly denser than average began to develop a greater density contrast with time because the gravitational forces were also slightly stronger than average in these regions. Eventually, after hundreds of millions of years, the overdense regions stopped expanding, turned around, and eventually collapsed to make bound objects such as galaxies. The gas within these collapsed objects cooled and fragmented into stars. This process, however, would have taken too long to explain the abundance of galaxies today, if it involved only the observed cosmic gas. Instead, gravity is strongly enhanced by the presence of dark matter – an unknown substance that makes up the vast majority (83%) of the cosmic density of matter. The motion of stars and gas around the centers of nearby galaxies indicates that each is surrounded by an extended mass of dark matter, and so dynamically-relaxed dark matter concentrations are generally referred to as "halos".

According to the standard cosmological model, the dark matter is cold (abbreviated as CDM), i.e., it behaves as a collection of collisionless particles that started out at matter domination with negligible thermal velocities and have evolved exclusively under gravitational forces. The model explains how both individual galaxies and the large-scale patterns in their distribution originated from the small initial density fluctuations. On the largest scales, observations of the present galaxy distribution have indeed found the same statistical patterns as seen in the CMB, enhanced as expected by billions of years of gravitational evolution. On smaller scales, the model describes how regions that were denser than average collapsed due to their enhanced gravity and eventually formed gravitationally-bound halos, first on small spatial scales and later on larger ones. In this hierarchical model of galaxy formation, the small galaxies formed first and then merged or accreted gas to form larger galaxies. At each snapshot of this cosmic evolution, the abundance of collapsed halos, whose masses are dominated by dark matter, can be computed from the initial conditions using numerical simulations. The common understanding of galaxy formation is based on the notion that stars formed out of the gas that cooled and subsequently condensed to high densities in the cores of some of these halos.

Gravity thus explains how some gas is pulled into the deep potential wells within dark matter halos and forms the galaxies. One might naively expect that the gas outside halos would remain mostly undisturbed. However, observations show that it has not remained neutral (i.e., in atomic form) but was largely ionized by the UV radiation emitted by the galaxies. The diffuse gas pervading the space outside and between galaxies is referred to as the intergalactic medium (IGM). For the first hundreds of millions of years after cosmological recombination, the so-called cosmic "dark ages", the universe was filled with diffuse atomic hydrogen. As soon as galaxies formed, they started to ionize diffuse hydrogen in their vicinity. Within less than a billion years, most of the IGM was re-ionized. We have not yet imaged the cosmic dark ages before the first galaxies had formed. One of the frontiers in current cosmological studies aims to study the cosmic epoch of reionization and the first generation of galaxies that triggered it.

1.2 The expanding universe

The modern physical description of the Universe as a whole can be traced back to Einstein, who assumed for simplicity the so-called "cosmological principle": that the distribution of matter and energy is homogeneous

and isotropic on the largest scales. Today isotropy is well established for the distribution of faint radio sources, optically-selected galaxies, the X-ray background, and most importantly the cosmic microwave background (hereafter, CMB). The constraints on homogeneity are less strict, but a cosmological model in which the Universe is isotropic but significantly inhomogeneous in spherical shells around our special location, is also excluded.

In General Relativity, the metric for a space-time which is spatially homogeneous and isotropic is the Friedman-Robertson-Walker metric, which can be written in the form

$$ds^2 = c^2 dt^2 - a^2(t) \left[\frac{dr^2}{1 - k\, r^2} + r^2 \left(d\theta^2 + \sin^2 \theta \, d\phi^2 \right) \right] \ , \tag{1}$$

where c is the speed of light, $a(t)$ is the cosmic scale factor which describes expansion in time t, and (r, θ, ϕ) are spherical comoving coordinates. The constant k determines the geometry of space; it is positive in a closed Universe, zero in a flat Universe (Euclidean space), and negative in an open Universe. Observers at rest remain at rest, at fixed (r, θ, ϕ), with their physical separation increasing with time in proportion to $a(t)$. A given observer sees a nearby observer at physical distance D receding at the Hubble velocity $H(t)D$, where the Hubble constant at time t is $H(t) = d\, a(t)/dt$. Light emitted by a source at time t is observed at $t = 0$ with a redshift $z = 1/a(t) - 1$, where we set $a(t = 0) \equiv 1$ for convenience.

The Einstein field equations of General Relativity yield the Friedmann equation

$$H^2(t) = \frac{8\pi G}{3} \rho - \frac{k}{a^2} \ , \tag{2}$$

which relates the expansion of the Universe to its matter-energy content. For each component of the energy density ρ, with an equation of state $p = p(\rho)$, the density ρ varies with $a(t)$ according to the thermodynamic relation

$$d(\rho c^2 r^3) = -p d(r^3) \ . \tag{3}$$

With the critical density

$$\rho_C(t) \equiv \frac{3H^2(t)}{8\pi G} \tag{4}$$

defined as the density needed for $k = 0$, we define the ratio of the total density to the critical density as

$$\Omega \equiv \frac{\rho}{\rho_C} \ . \tag{5}$$

With Ω_m, Ω_Λ, and Ω_r denoting the present contributions to Ω from matter (including cold dark matter as well as a contribution Ω_b from ordinary matter ["baryons"] made of protons and neutrons), vacuum energy (cosmological constant), and radiation, respectively, the Friedmann equation becomes

$$\frac{H(t)}{H_0} = \left[\frac{\Omega_m}{a^3} + \Omega_\Lambda + \frac{\Omega_r}{a^4} + \frac{\Omega_k}{a^2} \right] \ , \tag{6}$$

where we define H_0 and $\Omega_0 = \Omega_m + \Omega_\Lambda + \Omega_r$ to be the present values of H and Ω, respectively, and we let

$$\Omega_k \equiv -\frac{k}{H_0^2} = 1 - \Omega_m. \tag{7}$$

In the particularly simple Einstein-de Sitter model ($\Omega_m = 1$, $\Omega_\Lambda = \Omega_r = \Omega_k = 0$), the scale factor varies as $a(t) \propto t^{2/3}$. Even models with non-zero Ω_Λ or Ω_k approach the Einstein-de Sitter scaling-law at high redshift, i.e. when $(1 + z) \gg |{\Omega_m}^{-1} - 1|$ (as long as Ω_r can be neglected). In this high-z regime the age of the Universe is

$$t \approx \frac{2}{3H_0\sqrt{\Omega_m}} \, (1 + z)^{-3/2} \ \approx 10^9 \mathrm{yr} \left(\frac{1 + z}{7} \right)^{-3/2} . \tag{8}$$

Recent observations confine the standard set of cosmological parameters to a relatively narrow range. In particular, we seem to live in a universe dominated by a cosmological constant (Λ) and cold dark matter, or in short a ΛCDM cosmology (with Ω_k so small that it is usually assumed to equal zero) with an approximately scale-invariant primordial power spectrum of density fluctuations, i.e., $n \approx 1$ where the initial power spectrum is $P(k) = |\delta_{\mathbf{k}}|^2 \propto k^n$ in terms of the wavenumber k of the Fourier modes $\delta_{\mathbf{k}}$ (see §2.1 below). Also, the Hubble constant today is written as $H_0 = 100h$ km s^{-1}Mpc^{-1} in terms of h, and the overall normalization of the power spectrum is specified in terms of σ_8, the root-mean-square amplitude of mass fluctuations in spheres of radius 8 h^{-1} Mpc. For example, the best-fit cosmological parameters matching the WMAP data together with large-scale surveys of galaxies and supernovae are $\sigma_8 = 0.81$, $n = 0.96$, $h = 0.72$, $\Omega_m = 0.28$, $\Omega_\Lambda = 0.72$ and $\Omega_b = 0.046$.

2 Galaxy Formation

2.1 Growth of linear perturbations

As noted in the Introduction, observations of the CMB show that the universe at cosmic recombination (redshift $z \sim 10^3$) was remarkably uniform apart from spatial fluctuations in the energy density and in the gravitational potential of roughly one part in $\sim 10^5$. The primordial inhomogeneities in the density distribution grew over time and eventually led to the formation of galaxies as well as galaxy clusters and large-scale structure. In the early stages of this growth, as long as the density fluctuations on the relevant scales were much smaller than unity, their evolution can be understood with a linear perturbation analysis.

As before, we distinguish between fixed and comoving coordinates. Using vector notation, the fixed coordinate $\mathbf{r}$ corresponds to a comoving position $\mathbf{x} = \mathbf{r}/a$. In a homogeneous Universe with density ρ, we describe the cosmological expansion in terms of an ideal pressureless fluid of particles each of which is at fixed $\mathbf{x}$, expanding with the Hubble flow $\mathbf{v} = H(t)\mathbf{r}$ where $\mathbf{v} = d\mathbf{r}/dt$. Onto this uniform expansion we impose small perturbations, given by a relative density perturbation

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{r})}{\bar{\rho}} - 1 \ , \tag{9}$$

where the mean fluid density is $\bar{\rho}$, with a corresponding peculiar velocity $\mathbf{u} \equiv \mathbf{v} - H\mathbf{r}$. Then the fluid is described by the continuity and Euler equations in comoving coordinates:

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla \cdot [(1+\delta)\mathbf{u}] = 0 \tag{10}$$

$$\frac{\partial \mathbf{u}}{\partial t} + H\mathbf{u} + \frac{1}{a}(\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{a}\nabla\phi \ . \tag{11}$$

The potential ϕ is given by the Poisson equation, in terms of the density perturbation:

$$\nabla^2\phi = 4\pi G\bar{\rho}a^2\delta \ . \tag{12}$$

This fluid description is valid for describing the evolution of collisionless cold dark matter particles until different particle streams cross. This "shell-crossing" typically occurs only after perturbations have grown to become non-linear, and at that point the individual particle trajectories must in general be followed. Similarly, baryons can be described as a pressureless fluid as long as their temperature is negligibly small, but non-linear collapse leads to the formation of shocks in the gas.

For small perturbations $\delta \ll 1$, the fluid equations can be linearized and combined to yield

$$\frac{\partial^2\delta}{\partial t^2} + 2H\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta \ . \tag{13}$$

This linear equation has in general two independent solutions, only one of which grows with time. Starting with random initial conditions, this "growing mode" comes to dominate the density evolution. Thus, until it becomes non-linear, the density perturbation maintains its shape in comoving coordinates and grows in proportion to a growth factor $D(t)$. The growth factor in the matter-dominated era is given by

$$D(t) \propto \frac{\left(\Omega_\Lambda a^3 + \Omega_k a + \Omega_m\right)^{1/2}}{a^{3/2}} \int_0^a \frac{a'^{3/2}\,da'}{\left(\Omega_\Lambda a'^3 + \Omega_k a' + \Omega_m\right)^{3/2}} \ , \tag{14}$$

where we neglect Ω_r when considering halos forming in the matter-dominated regime at $z \ll 10^4$. In the Einstein-de Sitter model (or, at high redshift, in other models as well) the growth factor is simply proportional to $a(t)$.

The spatial form of the initial density fluctuations can be described in Fourier space, in terms of Fourier components

$$\delta_{\mathbf{k}} = \int d^3x\, \delta(x) e^{-i\mathbf{k}\cdot\mathbf{x}} \ . \tag{15}$$

Here we use the comoving wave-vector $\mathbf{k}$, whose magnitude k is the comoving wavenumber which is equal to 2π divided by the wavelength. The Fourier description is particularly simple for fluctuations generated by inflation. Inflation generates perturbations given by a Gaussian random field, in which different $\mathbf{k}$-modes are statistically independent, each with a random phase. The statistical properties of the fluctuations are determined by the variance of the different $\mathbf{k}$-modes, and the variance is described in terms of the power spectrum $P(k)$ as follows:

$$\langle \delta_{\mathbf{k}} \delta^*_{\mathbf{k}'} \rangle = (2\pi)^3 P(k)\, \delta^{(3)}\left(\mathbf{k} - \mathbf{k}'\right) \ , \tag{16}$$

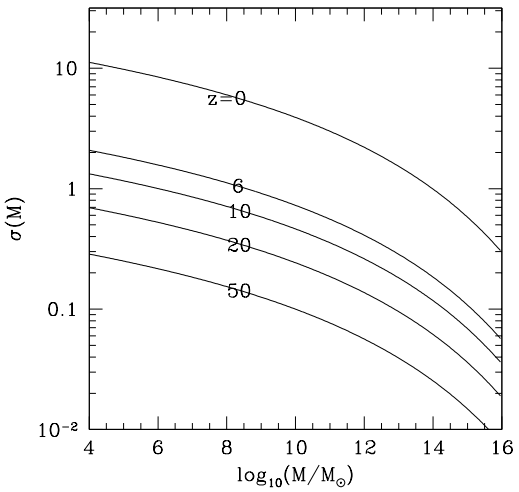

Figure 3: The root-mean-square amplitude of linearly-extrapolated density fluctuations σ as a function of mass M (in solar masses $M_\odot$, within a spherical top-hat filter) at different redshifts z. Halos form in regions that exceed the background density by a factor of order unity. This threshold is only surpassed by rare (many-σ) peaks for high masses at high redshifts. **Figure credit:** Loeb, A., *"How Did the First Stars and Galaxies Form?"*, Princeton University Press (2010).

where $\delta^{(3)}$ is the three-dimensional Dirac delta function. The gravitational potential fluctuations are sourced by the density fluctuations through Poisson's equation.

In standard models, inflation produces a primordial power-law spectrum $P(k) \propto k^n$ with $n \sim 1$. Perturbation growth in the radiation-dominated and then matter-dominated Universe results in a modified final power spectrum, characterized by a turnover at a scale of order the horizon cH^{-1} at matter-radiation equality, and a small-scale asymptotic shape of $P(k) \propto k^{n-4}$. The overall amplitude of the power spectrum is not specified by current models of inflation, and it is usually set by comparing to the observed CMB temperature fluctuations or to local measures of large-scale structure.

Since density fluctuations may exist on all scales, in order to determine the formation of objects of a given size or mass it is useful to consider the statistical distribution of the smoothed density field. Using a window function $W(\mathbf{r})$ normalized so that $\int d^3r\, W(\mathbf{r}) = 1$, the smoothed density perturbation field, $\int d^3r \delta(\mathbf{x}) W(\mathbf{r})$, itself follows a Gaussian distribution with zero mean. For the particular choice of a spherical top-hat, in which $W = 1$ in a sphere of radius R and is zero outside, the smoothed perturbation field measures the fluctuations in the mass in spheres of radius R. The normalization of the present power spectrum is often specified by the value of $\sigma_8 \equiv \sigma(R = 8h^{-1}\mathrm{Mpc})$. For the top-hat, the smoothed perturbation field is denoted δ_R or δ_M, where the mass M is related to the comoving radius R by $M = 4\pi\rho_m R^3/3$, in terms of the current mean density of matter ρ_m. The variance $\langle\delta_M\rangle^2$ is

$$\sigma^2(M) = \sigma^2(R) = \int_0^\infty \frac{dk}{2\pi^2}\, k^2 P(k) \left[\frac{3j_1(kR)}{kR}\right]^2 , \tag{17}$$

where $j_1(x) = (\sin x - x\cos x)/x^2$. The function $\sigma(M)$, plotted in Figure 3, plays a crucial role in estimates of the abundance of collapsed objects.

Different physical processes contributed to the perturbation growth. In the absence of other influences, gravitational forces due to density perturbations imprinted by inflation would have driven parallel perturbation growth in the dark matter, baryons and photons. However, since the photon sound speed is of order the speed of light, the radiation pressure produced sound waves on a scale of order the cosmic horizon and suppressed sub-horizon perturbations in the photon density. The baryonic pressure similarly suppressed perturbations in the gas below the (much smaller) so-called baryonic *Jeans* scale. Since the formation of hydrogen at recombination had decoupled the cosmic gas from its mechanical drag on the CMB, the baryons subsequently began to fall into the pre-existing gravitational potential wells of the dark matter.

Spatial fluctuations developed in the gas temperature as well as in the gas density. Both the baryons and the dark matter were affected on small scales by the temperature fluctuations through the gas pressure. Compton heating due to scattering of the residual free electrons (constituting a fraction $\sim 10^{-4}$) with the CMB photons remained effective, keeping the gas temperature fluctuations tied to the photon temperature fluctuations, even for a time after recombination. The growth of linear perturbations can be calculated with

the standard CMBFAST code (http://www.cmbfast.org), after a modification to account for the fact that the speed of sound of the gas also fluctuates spatially.

After recombination, two main drivers affect the baryon density and temperature fluctuations, namely, the thermalization with the CMB and the gravitational force that attracts the baryons to the dark matter potential wells. The density perturbations in all species grow together on scales where gravity is unopposed, outside the horizon (i.e., at $k < 0.01$ Mpc^{-1} at $z \sim 1000$). At $z = 1200$ the perturbations in the baryon-photon fluid oscillate as acoustic waves on scales of order the sound horizon ($k \sim 0.01$ Mpc^{-1}), while smaller-scale perturbations in both the photons and baryons are damped by photon diffusion and the drag of the diffusing photons on the baryons. On sufficiently small scales the power spectra of baryon density and temperature roughly assume the shape of the dark matter fluctuations (except for the gas-pressure cutoff at the very smallest scales), due to the effect of gravitational attraction on the baryon density and of the resulting adiabatic expansion on the gas temperature. After the mechanical coupling of the baryons to the photons ends at $z \sim 1000$, the baryon density perturbations gradually grow towards the dark matter perturbations because of gravity. Similarly, after the thermal coupling ends at $z \sim 200$, the baryon temperature fluctuations are driven by adiabatic expansion towards a value of 2/3 of the density fluctuations. By $z = 200$ the baryon infall into the dark matter potentials is well advanced and adiabatic expansion is becoming increasingly important in setting the baryon temperature.

2.2 Halo properties

The small density fluctuations evidenced in the CMB grow over time as described in the previous subsection, until the perturbation δ becomes of order unity, and the full non-linear gravitational problem must be considered. The dynamical collapse of a dark matter halo can be solved analytically only in cases of particular symmetry. If we consider a region which is much smaller than the horizon cH^{-1}, then the formation of a halo can be formulated as a problem in Newtonian gravity, in some cases with minor corrections coming from General Relativity. The simplest case is that of spherical symmetry, with an initial ($t = t_i \ll t_0$) top-hat of uniform overdensity δ_i inside a sphere of radius R. Although this model is restricted in its direct applicability, the results of spherical collapse have turned out to be surprisingly useful in understanding the properties and distribution of halos in models based on cold dark matter.

The collapse of a spherical top-hat perturbation is described by the Newtonian equation (with a correction for the cosmological constant)

$$\frac{d^2r}{dt^2} = H_0^2 \Omega_\Lambda \, r - \frac{GM}{r^2} \ , \tag{18}$$

where r is the radius in a fixed (not comoving) coordinate frame, H_0 is the present-day Hubble constant, M is the total mass enclosed within radius r, and the initial velocity field is given by the Hubble flow $dr/dt = H(t)r$. The enclosed δ grows initially as $\delta_L = \delta_i D(t)/D(t_i)$, in accordance with linear theory, but eventually δ grows above δ_L. If the mass shell at radius r is bound (i.e., if its total Newtonian energy is negative) then it reaches a radius of maximum expansion and subsequently collapses. As demonstrated in the previous section, at the moment when the top-hat collapses to a point, the overdensity predicted by linear theory is $\delta_L = 1.686$ in the Einstein-de Sitter model, with only a weak dependence on Ω_m and Ω_Λ. Thus a tophat collapses at redshift z if its linear overdensity extrapolated to the present day (also termed the critical density of collapse) is

$$\delta_{\rm crit}(z) = \frac{1.686}{D(z)} \ , \tag{19}$$

where we set $D(z = 0) = 1$.

Even a slight violation of the exact symmetry of the initial perturbation can prevent the tophat from collapsing to a point. Instead, the halo reaches a state of virial equilibrium by violent relaxation (phase mixing). Using the virial theorem $U = -2K$ to relate the potential energy U to the kinetic energy K in the final state (implying that the virial radius is half the turnaround radius - where the kinetic energy vanishes), the final overdensity relative to the critical density at the collapse redshift is $\Delta_c = 18\pi^2 \simeq 178$ in the Einstein-de Sitter model, modified in a Universe with $\Omega_m + \Omega_\Lambda = 1$ to the fitting formula

$$\Delta_c = 18\pi^2 + 82d - 39d^2 \ , \tag{20}$$

where $d \equiv \Omega_m^z - 1$ is evaluated at the collapse redshift, so that

$$\Omega_m^z = \frac{\Omega_m (1+z)^3}{\Omega_m (1+z)^3 + \Omega_\Lambda + \Omega_k (1+z)^2} \ . \tag{21}$$

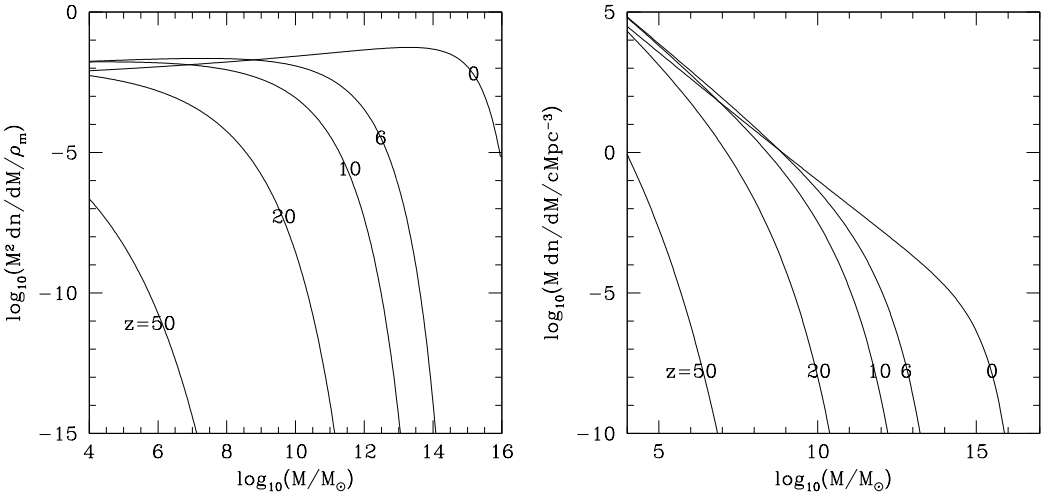

Figure 4: *Left panel:* The mass fraction incorporated into dark matter halos per logarithmic bin of halo mass $(M^2 dn/dM)/\rho_m$, as a function of M at different redshifts z. Here $\rho_m = \Omega_m \rho_c$ is the present-day matter density, and $n(M)dM$ is the comoving density of halos with masses between M and $M + dM$. The halo mass distribution was calculated based on an improved version of the Press-Schechter formalism for ellipsoidal collapse [Sheth, R. K., & Tormen, G. *Mon. Not. R. Astron. Soc.* **329**, 61 (2002)] that fits better numerical simulations. *Right panel:* Number density of halos per logarithmic bin of halo mass, Mdn/dM (in units of comoving Mpc^{-3}), at various redshifts. **Figure credit:** Loeb, A., *"How Did the First Stars and Galaxies Form?"*, Princeton University Press (2010).

A halo of mass M collapsing at redshift z thus has a virial radius

$$r_{\rm vir} = 0.784 \left(\frac{M}{10^8\ h^{-1}\ M_\odot}\right)^{1/3} \left[\frac{\Omega_m}{\Omega_m^z}\ \frac{\Delta_c}{18\pi^2}\right]^{-1/3} \left(\frac{1+z}{10}\right)^{-1}\ h^{-1}\ \mathrm{kpc}\ , \tag{22}$$

and a corresponding circular velocity,

$$V_c = \left(\frac{GM}{r_{\rm vir}}\right)^{1/2} = 23.4 \left(\frac{M}{10^8\ h^{-1}\ M_\odot}\right)^{1/3} \left[\frac{\Omega_m}{\Omega_m^z}\ \frac{\Delta_c}{18\pi^2}\right]^{1/6} \left(\frac{1+z}{10}\right)^{1/2}\ \mathrm{km\ s^{-1}}\ . \tag{23}$$

In these expressions we have assumed a present Hubble constant written in the form $H_0 = 100\,h$ km s^{-1}Mpc^{-1}. We may also define a virial temperature

$$T_{\rm vir} = \frac{\mu m_p V_c^2}{2k} = 1.98 \times 10^4\ \left(\frac{\mu}{0.6}\right) \left(\frac{M}{10^8\ h^{-1}\ M_\odot}\right)^{2/3} \left[\frac{\Omega_m}{\Omega_m^z}\ \frac{\Delta_c}{18\pi^2}\right]^{1/3} \left(\frac{1+z}{10}\right)\ \mathrm{K}\ , \tag{24}$$

where μ is the mean molecular weight and m_p is the proton mass. Note that the value of μ depends on the ionization fraction of the gas; for a fully ionized primordial gas $\mu = 0.59$, while a gas with ionized hydrogen but only singly-ionized helium has $\mu = 0.61$. The binding energy of the halo is approximately[1]

$$E_b = \frac{1}{2}\frac{GM^2}{r_{\rm vir}} = 5.45 \times 10^{53} \left(\frac{M}{10^8\ h^{-1}\ M_\odot}\right)^{5/3} \left[\frac{\Omega_m}{\Omega_m^z}\ \frac{\Delta_c}{18\pi^2}\right]^{1/3} \left(\frac{1+z}{10}\right) h^{-1}\ \mathrm{erg}\ . \tag{25}$$

Note that the binding energy of the baryons is smaller by a factor equal to the baryon fraction Ω_b/Ω_m.

Although spherical collapse captures some of the physics governing the formation of halos, structure formation in cold dark matter models proceeds hierarchically. At early times, most of the dark matter is in low-mass halos, and these halos continuously accrete and merge to form high-mass halos (see Figure 4). Numerical simulations of hierarchical halo formation indicate a roughly universal spherically-averaged density profile for the resulting halos, though with considerable scatter among different halos. The typical profile has the form

$$\rho(r) = \frac{3H_0^2}{8\pi G}(1+z)^3 \frac{\Omega_m}{\Omega_m^z} \frac{\delta_c}{c_{\rm N} x (1 + c_{\rm N} x)^2}\ , \tag{26}$$

[1]The coefficient of 1/2 in equation (25) would be exact for a singular isothermal sphere with $\rho(r) \propto 1/r^2$.

where $x = r/r_{\rm vir}$, and the characteristic density δ_c is related to the concentration parameter $c_{\rm N}$ by

$$\delta_c = \frac{\Delta_c}{3} \frac{c_{\rm N}^3}{\ln(1+c_{\rm N}) - c_{\rm N}/(1+c_{\rm N})} \ . \tag{27}$$

The concentration parameter itself depends on the halo mass M, at a given redshift z.

2.3 Formation of the first stars

Theoretical expectations for the properties of the first galaxies are based on the standard cosmological model outlined in the Introduction. The formation of the first bound objects marked the central milestone in the transition from the initial simplicity (discussed in the previous subsection) to the present-day complexity. Stars and accreting black holes output copious radiation and also produced explosions and outflows that brought into the IGM chemical products from stellar nucleosynthesis and enhanced magnetic fields. However, the formation of the very first stars, in a universe that had not yet suffered such feedback, remains a well-specified problem for theorists.

Stars form when large amounts of matter collapse to high densities. However, the process can be stopped if the pressure exerted by the hot intergalactic gas prevents outlying gas from falling into dark matter concentrations. As the gas falls into a dark matter halo, it forms shocks due to converging supersonic flows and in the process heats up and can only collapse further by first radiating its energy away. This restricts this process of collapse to very large clumps of dark matter that are around 100 000 times the mass of the Sun. Inside these clumps, the shocked gas loses energy by emitting radiation from excited molecular hydrogen that formed naturally within the primordial gas mixture of hydrogen and helium.

The first stars are expected to have been quite different from the stars that form today in the Milky Way. The higher pressure within the primordial gas due to the presence of fewer cooling agents suggests that fragmentation only occurred into relatively large units, in which gravity could overcome the pressure. Due to the lack of carbon, nitrogen, and oxygen – elements that would normally dominate the nuclear energy production in modern massive stars – the first stars must have condensed to extremely high densities and temperatures before nuclear reactions were able to heat the gas and balance gravity. These unusually massive stars produced high luminosities of UV photons, but their nuclear fuel was exhausted after 2–3 million years, resulting in a huge supernova or in collapse to a black hole. The heavy elements which were dispersed by the first supernovae in the surrounding gas, enabled the enriched gas to cool more effectively and fragment into lower mass stars. Simple calculations indicate that a carbon or oxygen enrichment of merely $<10^{-3}$ of the solar abundance is sufficient to allow solar mass stars to form. These second-generation "low-metallicity" stars are long-lived and could in principle be discovered in the halo of the Milky Way galaxy, providing fossil record of the earliest star formation episode in our cosmic environment.

Advances in computing power have made possible detailed numerical simulations of how the first stars formed. These simulations begin in the early universe, in which dark matter and gas are distributed uniformly, apart from tiny variations in density and temperature that are statistically distributed according to the patterns observed in the CMB. In order to span the vast range of scales needed to simulate an individual star within a cosmological context, the adopted codes zoom in repeatedly on the densest part of the first collapsing cloud that is found within the simulated volume. The simulation follows gravity, hydrodynamics, and chemical processes in the primordial gas, and resolves a scale that is > 10 orders of magnitudes smaller than that of the simulated box. In state-of-the-art simulations, the resolved scale is approaching the scale of the proto-star. The simulations have established that the first stars formed within halos containing $\sim 10^5 M_\odot$ in total mass, and indicate that the first stars most likely weighed tens to hundreds of solar masses each.

To estimate *when* the first stars formed we must remember that the first 100 000 solar mass halos collapsed in regions that happened to have a particularly high density enhancement very early on. There was initially only a small abundance of such regions in the entire universe, so a simulation that is limited to a small volume is unlikely to find such halos until much later. Simulating the entire universe is well beyond the capabilities of current simulations, but analytical models predict that the first observable star in the universe probably formed 30 million years after the Big Bang, less than a quarter of one percent of the Universe's total age of 13.7 billion years.

Although stars were extremely rare at first, gravitational collapse increased the abundance of galactic halos and star formation sites with time (Figure 2). Radiation from the first stars is expected to have eventually dissociated all the molecular hydrogen in the intergalactic medium, leading to the domination of a second generation of larger galaxies where the gas cooled via radiative transitions in atomic hydrogen and helium. Atomic cooling occurred in halos of mass above $\sim 10^8 M_\odot$, in which the infalling gas was heated above 10,000 K and became ionized. The first galaxies to form through atomic cooling are expected to have formed around redshift 45, and such galaxies were likely the main sites of star formation by the time reionization began in earnest. As the IGM was heated above 10,000 K by reionization, its pressure jumped

Figure 5: A full scale model of the James Webb Space Telescope (JWST), the successor to the Hubble Space Telescope. JWST includes a primary mirror 6.5 meters in diameter, and offers instrument sensitivity across the infrared wavelength range of 0.6–28μm which will allow detection of the first galaxies. The size of the Sun shield (the large flat screen in the image) is 22 meters×10 meters (72 ft×29 ft). The telescope will orbit 1.5 million kilometers from Earth at the Lagrange L2 point. **Image credit:** JWST/NASA (http://www.jwst.nasa.gov/).

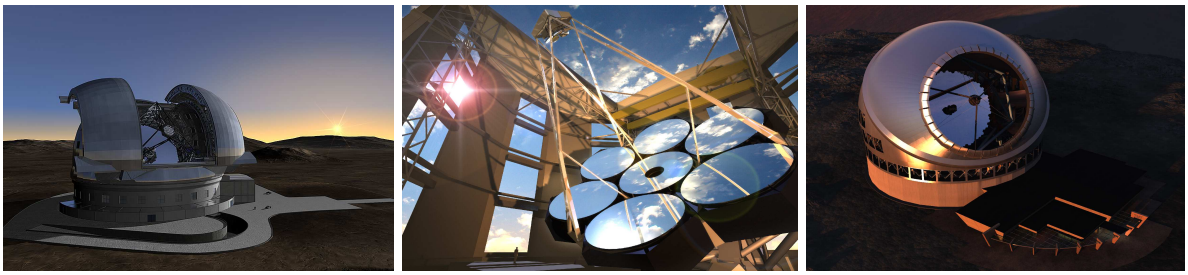

Figure 6: Artist's conception of the designs for three future giant telescopes that will be able to probe the first generation of galaxies from the ground: the European Extremely Large Telescope (EELT, left), the Giant Magellan Telescope (GMT, middle), and the Thirty Meter Telescope (TMT, right). **Image credits:** http://www.eso.org/sci/facilities/eelt/, http://www.gmto.org/, and http://www.tmt.org/.

and prevented the gas from accreting into newly forming halos below $\sim 10^9 M_\odot$. The first Milky-Way-sized halo $M = 10^{12} M_\odot$ is predicted to have formed 400 million years after the Big Bang, but such halos have become typical galactic hosts only in the last five billion years.

Hydrogen is the most abundant element in the Universe, The prominent Lyman-α spectral line of hydrogen (corresponding to a transition from its first excited level to its ground state) provides an important probe of the condensation of primordial gas into the first galaxies. Existing searches for Lyman-α emission have discovered galaxies robustly out to a redshift $z \sim 7$ with some unconfirmed candidate galaxies out to $z \sim 10$. The spectral break owing to Lyman-α absorption by the IGM allows to identify high-redshifts galaxies photometrically. Existing observations provide only a preliminary glimpse into the formation of the first galaxies.

Within the next decade, NASA plans to launch an infrared space telescope (*JWST*; Figure 5) that will image some of the earliest sources of light (stars and black holes) in the Universe. In parallel, there are several initiatives to construct large-aperture infrared telescopes on the ground with the same goal in mind.

The next generation of ground-based telescopes will have a diameter of twenty to thirty meters (Figure 6). Together with *JWST* (which will not be affected by the atmospheric background) they will be able to image and make spectral studies of the early galaxies. Given that these galaxies also create the ionized bubbles around them by their UV emission, during reionization the locations of galaxies should correlate with bubbles within the neutral hydrogen. Within a decade it should be possible to explore the environmental influence of individual galaxies by using these telescopes in combination with 21-cm probes of reionization.

2.4 Gamma-ray Bursts: probing the first stars one star at a time

So far, to learn about diffuse IGM gas pervading the space outside and between galaxies, astronomers routinely study its absorption signatures in the spectra of distant quasars, the brightest long-lived astronomical objects. Quasars' great luminosities are believed to be powered by accretion of gas onto black holes weighing

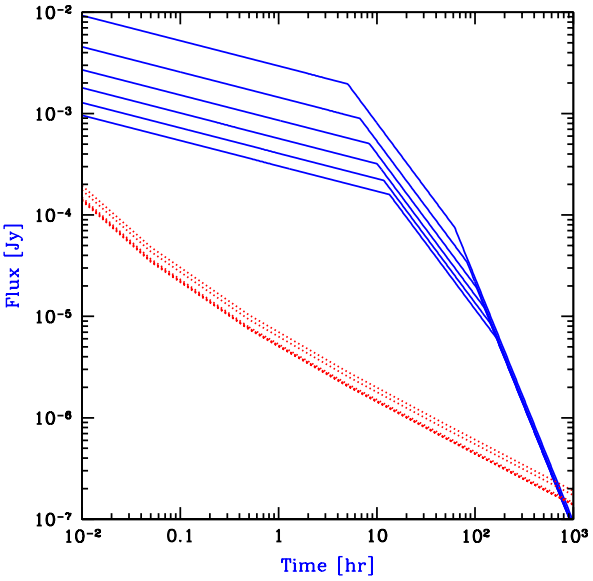

Figure 7: GRB afterglow flux as a function of time since the γ-ray trigger in the observer frame. The flux (solid curves) is calculated at the redshifted Lyman-α wavelength. The dotted curves show the planned detection threshold for the *James Webb Space Telescope* (*JWST*), assuming a spectral resolution $R = 5000$ with the near infrared spectrometer, a signal to noise ratio of 5 per spectral resolution element, and an exposure time equal to 20% of the time since the GRB explosion. Each set of curves shows a sequence of redshifts, namely $z = 5$, 7, 9, 11, 13, and 15, respectively, from top to bottom. **Figure credit:** Barkana, R., & Loeb, A., *Astrophys. J.* **601, 64 (2004).**

up to a few billion times the mass of the Sun that are situated in the centers of massive galaxies. As the surrounding gas spirals in toward the black hole sink, the viscous dissipation of heat makes the gas glow brightly into space, creating a luminous source visible from afar.

Over the past decade, an alternative population of bright sources at cosmological distances was discovered, the so-called afterglows of *Gamma-Ray Bursts* (GRBs). These events are characterized by a flash of high-energy (> 0.1 MeV) photons, typically lasting 0.1–100 seconds, which is followed by an afterglow of lower-energy photons over much longer timescales. The afterglow peaks at X-ray, UV, optical and eventually radio wavelengths on time scales of minutes, hours, days, and months, respectively. The central engines of GRBs are believed to be associated with the compact remnants (neutron stars or stellar-mass black holes) of massive stars. Their high luminosities make them detectable out to the edge of the visible Universe. GRBs offer the opportunity to detect the most distant (and hence earliest) population of massive stars, the so-called Population III (or Pop III), one star at a time. In the hierarchical assembly process of halos that are dominated by cold dark matter (CDM), the first galaxies should have had lower masses (and lower stellar luminosities) than their more recent counterparts. Consequently, the characteristic luminosity of galaxies or quasars is expected to decline with increasing redshift. GRB afterglows, which already produce a peak flux comparable to that of quasars or starburst galaxies at $z \sim 1-2$, are therefore expected to outshine any competing source at the highest redshifts, when the first dwarf galaxies formed in the Universe.

GRBs, the electromagnetically-brightest explosions in the Universe, should be detectable out to redshifts $z > 10$. High-redshift GRBs can be identified through infrared photometry, based on the Lyman-α break induced by absorption of their spectrum at wavelengths below $1.216\,\mu\mathrm{m}\,[(1+z)/10]$. Follow-up spectroscopy of high-redshift candidates can then be performed on a 10-meter-class telescope. GRB afterglows offer the opportunity to detect stars as well as to probe the metal enrichment level of the intervening IGM. Recently, the *Swift* satellite has detected a GRB originating at $z \simeq 8.3$, thus demonstrating the viability of GRBs as probes of the early Universe.

Another advantage of GRBs is that the GRB afterglow flux at a given observed time lag after the γ-ray trigger is not expected to fade significantly with increasing redshift, since higher redshifts translate to earlier times in the source frame, during which the afterglow is intrinsically brighter. For standard afterglow lightcurves and spectra, the increase in the luminosity distance with redshift is compensated by this cosmological time-stretching effect as shown in Figure 7.

GRB afterglows have smooth (broken power-law) continuum spectra unlike quasars which show strong spectral features (such as broad emission lines or the so-called "blue bump") that complicate the extraction

of IGM absorption features. In particular, the extrapolation into the spectral regime marked by the IGM Lyman-α absorption during the epoch of reionization is much more straightforward for the smooth UV spectra of GRB afterglows than for quasars with an underlying broad Lyman-α emission line. However, the interpretation may be complicated by the presence of damped Lyman-α absorption by dense neutral hydrogen in the immediate environment of the GRB within its host galaxy. Since GRBs originate from the dense environment of active star formation, such damped absorption is expected and indeed has been seen, including in the most distant GRB at $z = 8.3$.

2.5 Supermassive black holes

The fossil record in the present-day Universe indicates that every bulged galaxy hosts a supermassive black hole (BH) at its center. This conclusion is derived from a variety of techniques which probe the dynamics of stars and gas in galactic nuclei. The inferred BHs are dormant or faint most of the time, but occasionally flash in a short burst of radiation that lasts for a small fraction of the age of the Universe. The short duty cycle accounts for the fact that bright quasars are much less abundant than their host galaxies, but it begs the more fundamental question: *why is the quasar activity so brief?* A natural explanation is that quasars are suicidal, namely the energy output from the BHs regulates their own growth.

Supermassive BHs make up a small fraction, $< 10^{-3}$, of the total mass in their host galaxies, and so their direct dynamical impact is limited to the central star distribution where their gravitational influence dominates. Dynamical friction on the background stars keeps the BH close to the center. Random fluctuations in the distribution of stars induces a Brownian motion of the BH. This motion can be described by the same Langevin equation that captures the motion of a massive dust particle as it responds to random kicks from the much lighter molecules of air around it. The characteristic speed by which the BH wanders around the center is small, $\sim (m_\star/M_{\rm BH})^{1/2}\sigma_\star$, where $m_\star$ and $M_{\rm BH}$ are the masses of a single star and the BH, respectively, and $\sigma_\star$ is the stellar velocity dispersion. Since the random force fluctuates on a dynamical time, the BH wanders across a region that is smaller by a factor of $\sim (m_\star/M_{\rm BH})^{1/2}$ than the region traversed by the stars inducing the fluctuating force on it.

The dynamical insignificance of the BH on the global galactic scale is misleading. The gravitational binding energy per rest-mass energy of galaxies is of order $\sim (\sigma_\star/c)^2 < 10^{-6}$. Since BH are relativistic objects, the gravitational binding energy of material that feeds them amounts to a substantial fraction its rest mass energy. Even if the BH mass amounts to a fraction as small as $\sim 10^{-4}$ of the baryonic mass in a galaxy, and only a percent of the accreted rest-mass energy is deposited into the gaseous environment of the BH, this slight deposition can unbind the entire gas reservoir of the host galaxy. This order-of-magnitude estimate explains why quasars may be short lived. As soon as the central BH accretes large quantities of gas so as to significantly increase its mass, it releases large amounts of energy and momentum that could suppress further accretion onto it. In short, the BH growth might be *self-regulated*.

The principle of *self-regulation* naturally leads to a correlation between the final BH mass, $M_{\rm bh}$, and the depth of the gravitational potential well to which the surrounding gas is confined. The latter can be characterized by the velocity dispersion of the associated stars, $\sim \sigma_\star^2$. Indeed a correlation between $M_{\rm bh}$ and $\sigma_\star^4$ is observed in the present-day Universe. If quasars shine near their Eddington limit as suggested by observations of low and high-redshift quasars, then a fraction of $\sim$ 5–10% of the energy released by the quasar over a galactic dynamical time needs to be captured in the surrounding galactic gas in order for the BH growth to be self-regulated. With this interpretation, the $M_{\rm bh}$–$\sigma_\star$ relation reflects the limit introduced to the BH mass by self-regulation; deviations from this relation are inevitable during episodes of BH growth or as a result of mergers of galaxies that have no cold gas in them. A physical scatter around this upper envelope could also result from variations in the efficiency by which the released BH energy couples to the surrounding gas.

Various prescriptions for self-regulation were sketched in the literature. These involve either energy or momentum-driven winds, with the latter type being a factor of $\sim v_c/c$ less efficient. The quasar remains active during the dynamical time of the initial gas reservoir, $\sim 10^7$ years, and fades afterwards due to the dilution of this reservoir. The BH growth may resume if the cold gas reservoir is replenished through a new merger. Following early analytic work, extensive numerical simulations demonstrated that galaxy mergers do produce the observed correlations between black hole mass and spheroid properties. Because of the limited resolution near the galaxy nucleus, these simulations adopt a simple prescription for the accretion flow that feeds the black hole. The actual feedback in reality may depend crucially on the geometry of this flow and the physical mechanism that couples the energy or momentum output of the quasar to the surrounding gas.

The inflow of cold gas towards galaxy centers during the growth phase of the BH would naturally be accompanied by a burst of star formation. The fraction of gas that is not consumed by stars or ejected by supernova-driven winds, will continue to feed the BH. It is therefore not surprising that quasar and starburst activities co-exist in Ultra Luminous Infrared Galaxies, and that all quasars show broad metal

lines indicating pre-enrichment of the surrounding gas with heavy elements.

The upper mass of galaxies may also be regulated by the energy output from quasar activity. This would account for the fact that cooling flows are suppressed in present-day X-ray clusters, and that massive BHs and stars in galactic bulges were already formed at $z \sim 2$. In the cores of cooling X-ray clusters, there is often an active central BH that supplies sufficient energy to compensate for the cooling of the gas. The primary physical process by which this energy couples to the gas is still unknown.

The quasars discovered so far at $z \sim 6$ mark the early growth of the most massive BHs and galactic spheroids. The BHs powering these bright quasars possess a mass of a few billion solar masses. A quasar radiating at its Eddington limiting luminosity, $L_E = 1.4 \times 10^{47}$ erg s$^{-1}(M_{\rm bh}/10^9 M_\odot)$, with a radiative efficiency, $\epsilon_{\rm rad} = L_E/\dot{M}c^2$, for converting accreted mass into radiation, would grow exponentially in mass as a function of time t, $M_{\rm bh} = M_{\rm seed}\exp\{t/t_E\}$ from its initial seed mass $M_{\rm seed}$, on a time scale, $t_E = 4.1 \times 10^7$ yr$(\epsilon_{\rm rad}/0.1)$. Thus, the required growth time in units of the Hubble time $t_{\rm hubble} = 10^9$ yr$[(1+z)/7]^{-3/2}$ is

$$\frac{t_{\rm growth}}{t_{\rm hubble}} = 0.7\left(\frac{\epsilon_{\rm rad}}{10\%}\right)\left(\frac{1+z}{7}\right)^{3/2}\ln\left(\frac{M_{\rm bh}/10^9 M_\odot}{M_{\rm seed}/100 M_\odot}\right) \ . \tag{28}$$

The age of the Universe at $z \sim 6$ provides just sufficient time to grow a BH with $M_{\rm bh} \sim 10^9 M_\odot$ out of a stellar mass seed with $\epsilon_{\rm rad} = 10\%$. The growth time is shorter for smaller radiative efficiencies or a higher seed mass.

2.6 The epoch of reionization

Given the understanding described above of how many galaxies formed at various times, the course of reionization can be determined universe-wide by counting photons from all sources of light. Both stars and black holes contribute ionizing photons, but the early universe is dominated by small galaxies which in the local universe have central black holes that are disproportionately small, and indeed quasars are rare above redshift 6. Thus, stars most likely dominated the production of ionizing UV photons during the reionization epoch [although high-redshift galaxies should have also emitted X-rays from accreting black holes and accelerated particles in collisionless shocks]. Since most stellar ionizing photons are only slightly more energetic than the 13.6 eV ionization threshold of hydrogen, they are absorbed efficiently once they reach a region with substantial neutral hydrogen). This makes the IGM during reionization a two-phase medium characterized by highly ionized regions separated from neutral regions by sharp ionization fronts.

We can obtain a first estimate of the requirements of reionization by demanding one stellar ionizing photon for each hydrogen atom in the IGM. If we conservatively assume that stars within the early galaxies were similar to those observed locally, then each star produced ~ 4000 ionizing photons per baryon. Star formation is observed today to be an inefficient process, but even if stars in galaxies formed out of only $\sim 10\%$ of the available gas, it was still sufficient to accumulate a small fraction (of order 0.1%) of the total baryonic mass in the universe into galaxies in order to ionize the entire IGM. More accurate estimates of the actual required fraction account for the formation of some primordial stars (which were massive, efficient ionizers, as discussed above), and for recombinations of hydrogen atoms at high redshifts and in dense regions.

From studies of quasar absorption lines at $z \sim 6$ we know that the IGM is highly ionized a billion years after the big bang. There are hints, however, that some large neutral hydrogen regions persist at these early times and so this suggests that we may not need to go to much higher redshifts to begin to see the epoch of reionization. We now know that the universe could not have fully reionized earlier than an age of 300 million years, since WMAP observed the effect of the freshly created plasma at reionization on the large-scale polarization anisotropies of the CMB and this limits the reionization redshift; an earlier reionization, when the universe was denser, would have created a stronger scattering signature that would be inconsistent with the WMAP observations. In any case, the redshift at which reionization ended only constrains the overall cosmic efficiency of ionizing photon production. In comparison, a detailed picture of reionization as it happens will teach us a great deal about the population of young galaxies that produced this cosmic phase transition. A key point is that the spatial distribution of ionized bubbles is determined by clustered groups of galaxies and not by individual galaxies. At such early times galaxies were strongly clustered even on very large scales (up to tens of Mpc), and these scales therefore dominate the structure of reionization. The basic idea is simple. At high redshift, galactic halos are rare and correspond to rare, high density peaks. As an analogy, imagine searching on Earth for mountain peaks above 5000 meters. The 200 such peaks are not at all distributed uniformly but instead are found in a few distinct clusters on top of large mountain ranges. Given the large-scale boost provided by a mountain range, a small-scale crest need only provide a small additional rise in order to become a 5000 meter peak. The same crest, if it formed within a valley, would not come anywhere near 5000 meters in total height. Similarly, in order to find the early galaxies, one must first locate a region with a large-scale density enhancement, and then galaxies will be found there in abundance.

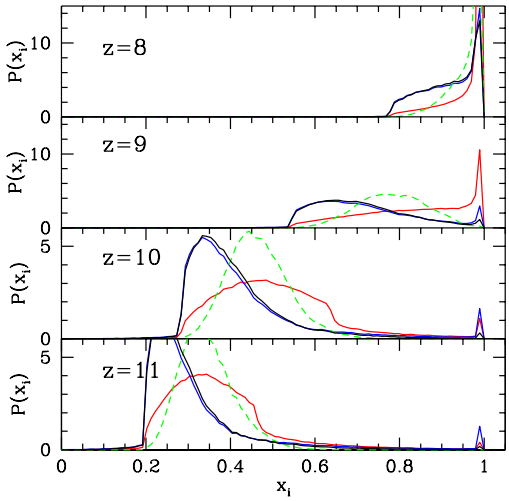

Figure 8: Probability distribution of x_i at redshifts $z = 8$, 9, 10, and 11, based on existing data on the CMB polarization anisotropies and the Lyman-α forest. **Figure credit:** Pritchard, J. R., Loeb, A.. & Wyithe, S., *Mon. Not. R. Astr. Soc.*, in press (2010); http://arxiv.org/abs/0908.3891.

The ionizing radiation emitted from the stars in each galaxy initially produces an isolated ionized bubble. However, in a region dense with galaxies the bubbles quickly overlap into one large bubble, completing reionization in this region while the rest of the universe is still mostly neutral. Most importantly, since the abundance of rare density peaks is very sensitive to small changes in the density threshold, even a large-scale region with a small enhanced density (say, 10% above the mean density of the universe) can have a much larger concentration of galaxies than in other regions (e.g., a 50% enhancement). On the other hand, reionization is harder to achieve in dense regions, since the protons and electrons collide and recombine more often in such regions, and newly-formed hydrogen atoms need to be reionized again by additional ionizing photons. However, the overdense regions end up reionizing first since the number of ionizing sources in these regions is increased so strongly. The large-scale topology of reionization is therefore inside out, with underdense voids reionizing only at the very end of reionization, with the help of extra ionizing photons coming in from their surroundings (which have a higher density of galaxies than the voids themselves). This is a key prediction awaiting observational testing.

Detailed analytical models that account for large-scale variations in the abundance of galaxies confirm that the typical bubble size starts well below a Mpc early in reionization, as expected for an individual galaxy, rises to 5–10 Mpc during the central phase (i.e., when the universe is half ionized), and then by another factor of $\sim$5 towards the end of reionization. These scales are given in comoving units that scale with the expansion of the universe, so that the actual sizes at a redshift z were smaller than these numbers by a factor of $(1 + z)$. Numerical simulations have only recently begun to reach the enormous scales needed to capture this evolution. Accounting precisely for gravitational evolution on a wide range of scales but still crudely for gas dynamics, star formation, and the radiative transfer of ionizing photons, the simulations confirm that the large-scale topology of reionization is inside out, and that this topology can be used to study the abundance and clustering of the ionizing sources.

The characteristic observable size of the ionized bubbles at the end of reionization can be calculated based on simple considerations that only depend on the power-spectrum of density fluctuations and the redshift. As the size of an ionized bubble increases, the time it takes a 21-cm photon emitted by hydrogen to traverse it gets longer. At the same time, the variation in the time at which different regions reionize becomes smaller as the regions grow larger. Thus, there is a maximum size above which the photon crossing time is longer than the cosmic variance in ionization time. Regions bigger than this size will be ionized at their near side by the time a 21-cm photon will cross them towards the observer from their far side. They would appear to the observer as one-sided, and hence signal the end of reionization. These considerations imply a characteristic size for the ionized bubbles of $\sim$ 10 physical Mpc at $z \sim 6$ (equivalent to 70 Mpc today). This result implies that future radio experiments should be tuned to a characteristic angular scale of tens of arcminutes for an optimal detection of 21-cm brightness fluctuations near the end of reionization (see §3.2).

Existing data on the polarization anisotropies of the CMB as well as the Lyman-α forest can be used to derive a probability distribution for the hydrogen ionization fraction (x_i) as a function of redshift. Figure 8 shows this likelihood distribution in four redshift bins of interest to upcoming observations. Although there

is considerable uncertainty in x_i at each redshift, it is evident from existing data that hydrogen is highly ionized by $z = 8$ (at least to $x_i > 0.8$).

To produce one ionizing photon per baryon requires a minimum comoving density of Milky-Way (so-called Population II) stars of,

$$\rho_\star \approx 1.7 \times 10^6 f_{\rm esc}^{-1}\ M_\odot\ {\rm Mpc}^{-3}, \tag{29}$$

or equivalently, a cosmological density parameter in stars of $\Omega_\star \sim 1.25 \times 10^{-5} f_{\rm esc}^{-1}$. More typically, the threshold for reionization involves at least a few ionizing photons per proton (with the right-hand-side being $\sim 10^{-6}$ cm^{-3}), since the recombination time at the mean density is comparable to the age of the Universe at $z \sim 10$.

For the local mass function of (Population II) stars at solar metallicity, the star formation rate per unit comoving volume that is required for balancing recombinations in an already ionized IGM, is given by

$$\dot{\rho_\star} \approx 2 \times 10^{-3} f_{\rm esc}^{-1} C \left(\frac{1+z}{10}\right)^3\ M_\odot\ {\rm yr}^{-1}\ {\rm Mpc}^{-3}, \tag{30}$$

where $C = \langle n_e^2 \rangle / \langle n_H \rangle^2$ is the volume-averaged clumpiness factor of the electron density up to some threshold overdensity of gas which remains neutral. Current state-of-the-art surveys (HST WFC3/IR) are only sensitive to the bright end of the luminosity function of galaxies at $z > 6$ and hence provide a lower limit on the production rate of ionizing photons during reionization.

2.7 Post-reionization suppression of low-mass galaxies

After the ionized bubbles overlapped in each region, the ionizing background increased sharply, and the IGM was heated by the ionizing radiation to a temperature $T_{\rm IGM} > 10^4$ K. Due to the substantial increase in the IGM pressure, the smallest mass scale into which the cosmic gas could fragment, the so-called Jeans mass, increased dramatically, changing the minimum mass of forming galaxies.

Gas infall depends sensitively on the Jeans mass. When a halo more massive than the Jeans mass begins to form, the gravity of its dark matter overcomes the gas pressure. Even in halos below the Jeans mass, although the gas is initially held up by pressure, once the dark matter collapses its increased gravity pulls in some gas. Thus, the Jeans mass is generally higher than the actual limiting mass for accretion. Before reionization, the IGM is cold and neutral, and the Jeans mass plays a secondary role in limiting galaxy formation compared to cooling. After reionization, the Jeans mass is increased by several orders of magnitude due to the photoionization heating of the IGM, and hence begins to play a dominant role in limiting the formation of stars. Gas infall in a reionized and heated Universe has been investigated in a number of numerical simulations. Three dimensional numerical simulations found a significant suppression of gas infall in even larger halos ($V_c \sim 75$ km s^{-1}), but this was mostly due to a suppression of late infall at z<2.

When a volume of the IGM is ionized by stars, the gas is heated to a temperature $T_{\rm IGM} \sim 10^4$ K. If quasars dominate the UV background at reionization, their harder photon spectrum leads to $T_{\rm IGM} > 2 \times 10^4$ K. Including the effects of dark matter, a given temperature results in a linear Jeans mass corresponding to a halo circular velocity of

$$V_J \approx 80 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4 {\rm K}}\right)^{1/2}\ {\rm km\ s}^{-1}. \tag{31}$$

In halos with a circular velocity well above V_J, the gas fraction in infalling gas equals the universal mean of Ω_b/Ω_m, but gas infall is suppressed in smaller halos. A simple estimate of the limiting circular velocity, below which halos have essentially no gas infall, is obtained by substituting the virial overdensity for the mean density in the definition of the Jeans mass. The resulting estimate is

$$V_{\rm lim} = 34 \left(\frac{T_{\rm IGM}}{1.5 \times 10^4 {\rm K}}\right)^{1/2}\ {\rm km\ s}^{-1}. \tag{32}$$

This value is in rough agreement with the numerical simulations mentioned before.

Although the Jeans mass is closely related to the rate of gas infall at a given time, it does not directly yield the total gas residing in halos at a given time. The latter quantity depends on the entire history of gas accretion onto halos, as well as on the merger histories of halos, and an accurate description must involve a time-averaged Jeans mass. The gas content of halos in simulations is well fit by an expression which depends on the filtering mass, a particular time-averaged Jeans mass.

The reionization process was not perfectly synchronized throughout the Universe. Large-scale regions with a higher density than the mean tended to form galaxies first and reionized earlier than underdense regions. The suppression of low-mass galaxies by reionization is therefore modulated by the fluctuations

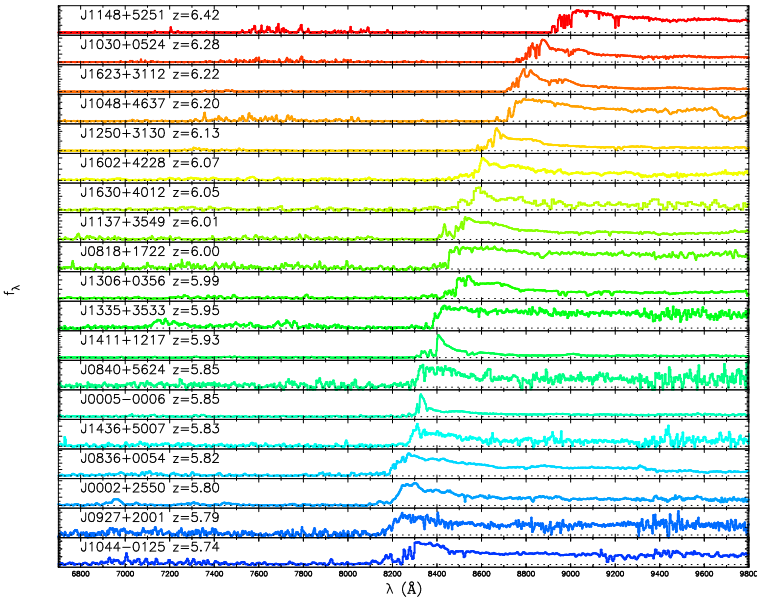

Figure 9: Spectra of 19 quasars with redshifts $5.74 < z < 6.42$ from the *Sloan Digital Sky Survey*. For some of the highest-redshift quasars, the spectrum shows no transmitted flux shortward of the Lyman-α wavelength at the quasar redshift (the so-called "Gunn-Peterson trough"), indicating a non-negligible neutral fraction in the IGM. **Figure credit:** Fan, X., et al. *Astron. J.* **125, 1649 (2005).**

in the timing of reionization. Inhomogeneous reionization imprint a signature on the power-spectrum of low-mass galaxies. Future high-redshift galaxy surveys hoping to constrain inflationary parameters must properly model the effects of reionization; conversely, they will also place new constraints on the thermal history of the IGM during reionization.

3 Probing the Diffuse Intergalactic Hydrogen

3.1 Lyman-alpha absorption

Resonant Lyman-α absorption has thus far proved to be the best probe of the state of the IGM. The optical depth to absorption by a uniform intergalactic medium is

$$\begin{aligned} \tau_s &= \frac{\pi e^2 f_\alpha \lambda_\alpha n_{\rm H\,I}\,(z)}{m_e c H(z)} \\ &\approx 6.45\times10^5 x_{\rm H\,I}\left(\frac{\Omega_b h}{0.0315}\right)\left(\frac{\Omega_m}{0.3}\right)^{-1/2}\left(\frac{1+z}{10}\right)^{3/2}, \end{aligned} \tag{33}$$

where $H \approx 100h$ km s^{-1} Mpc$^{-1}\Omega_m^{1/2}(1+z)^{3/2}$ is the Hubble parameter at redshift z; $f_\alpha = 0.4162$ and $\lambda_\alpha = 1216$Å are the oscillator strength and the wavelength of the Lyman-α transition; $n_{\rm H\,I}\,(z)$ is the neutral hydrogen density at z (assuming primordial abundances); Ω_m and Ω_b are the present-day density parameters of all matter and of baryons, respectively; and $x_{\rm H\,I}$ is the average fraction of neutral hydrogen. In the second equality we have implicitly considered high redshifts.

Lyman-α absorption is thus highly sensitive to the presence of even trace amounts of neutral hydrogen. The lack of full absorption in quasar spectra then implies that the IGM has been very highly ionized during much of the history of the universe, from at most a billion years after the big bang to the present time. At redshifts approaching six, however, the optical depth increases, and the observed absorption becomes very strong. The difference between the unabsorbed expectation and the actual observed spectrum can be used to measure the amount of absorption, and thus to infer the atomic hydrogen density.

Several quasars beyond $z \sim 6.1$ show in their spectra a strong (so-called "Gunn-Peterson") trough, a blank spectral region at wavelengths shorter than Lyα at the quasar redshift (Figure 9). The detection of Gunn-Peterson troughs indicates a rapid change in the neutral content of the IGM at $z \sim 6$, and hence a rapid change in the intensity of the background ionizing flux. However, even a small atomic hydrogen fraction of $\sim 10^{-3}$ would still produce nearly complete Lyα absorption.

While only resonant Lyα absorption is important at moderate redshifts, the damping wing of the Lyα line plays a significant role when neutral fractions of order unity are considered at $z > 6$. The scattering cross-section of the Lyα resonance line by neutral hydrogen is given by

$$\sigma_\alpha(\nu) = \frac{3\lambda_\alpha^2\Lambda_\alpha^2}{8\pi}\frac{(\nu/\nu_\alpha)^4}{4\pi^2(\nu-\nu_\alpha)^2 + (\Lambda_\alpha^2/4)(\nu/\nu_\alpha)^6} \ , \tag{34}$$

where $\Lambda_\alpha = (8\pi^2 e^2 f_\alpha/3m_e c\lambda_\alpha^2) = 6.25 \times 10^8$ s^{-1} is the Lyα ($2p \rightarrow 1s$) decay rate, $f_\alpha = 0.4162$ is the oscillator strength, and $\lambda_\alpha = 1216$Å and $\nu_\alpha = (c/\lambda_\alpha) = 2.47 \times 10^{15}$ Hz are the wavelength and frequency of the Lyα line. The term in the numerator is responsible for the classical Rayleigh scattering.

Although reionization is an inhomogeneous process, we consider here a simple illustrative case of instantaneous reionization. Consider a source at a redshift z_s beyond the redshift of reionization, $z_{\rm reion}$, and the corresponding scattering optical depth of a uniform, neutral IGM of hydrogen density $n_{\rm H,0}(1+z)^3$ between the source and the reionization redshift. The optical depth is a function of the observed wavelength $\lambda_{\rm obs}$,

$$\tau(\lambda_{\rm obs}) = \int_{z_{\rm reion}}^{z_s} dz\,\frac{cdt}{dz}\,n_{\rm H,0}(1+z)^3\sigma_\alpha\left[\nu_{\rm obs}(1+z)\right] \ , \tag{35}$$

where $\nu_{\rm obs} = c/\lambda_{\rm obs}$ and for a flat Universe with $(\Omega_m + \Omega_\Lambda) = 1$,

$$\frac{dt}{dz} = \left[(1+z)H(z)\right]^{-1} = H_0^{-1} \times \left[\Omega_m(1+z)^5 + \Omega_\Lambda(1+z)^2\right]^{-1/2} \ . \tag{36}$$

At wavelengths longer than Lyα at the source, the optical depth obtains a small value; these photons redshift away from the line center along its red wing and never resonate with the line core on their way to the observer. Considering only the regime in which $|\nu - \nu_\alpha| \gg \Lambda_\alpha$, we may ignore the second term in the denominator of equation (34). This leads to an analytical result for the red damping wing of the Gunn-Peterson trough,

$$\tau(\lambda_{\rm obs}) = \tau_s\left(\frac{\Lambda}{4\pi^2\nu_\alpha}\right)\tilde{\lambda}_{\rm obs}^{3/2}\left[I(\tilde{\lambda}_{\rm obs}^{-1}) - I([(1+z_{\rm reion})/(1+z_s)]\tilde{\lambda}_{\rm obs}^{-1})\right] \ , \tag{37}$$

an expression valid for $\tilde{\lambda}_{\rm obs} \geq 1$, where τ_s is given in equation (33), and we also define

$$\tilde{\lambda}_{\rm obs} \equiv \frac{\lambda_{\rm obs}}{(1+z_s)\lambda_\alpha} \tag{38}$$

and

$$I(x) \equiv \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\ln\left[\frac{1+x^{1/2}}{1-x^{1/2}}\right] \ . \tag{39}$$

3.2 21-cm absorption or emission

3.2.1 The spin temperature of the 21-cm transition of hydrogen

The ground state of hydrogen exhibits hyperfine splitting owing to the possibility of two relative alignments of the spins of the proton and the electron. The state with parallel spins (the triplet state) has a slightly higher energy than the state with anti-parallel spins (the singlet state). The 21-cm line associated with the spin-flip transition from the triplet to the singlet state is often used to detect neutral hydrogen in the local universe. At high redshift, the occurrence of a neutral pre-reionization IGM offers the prospect of detecting the first sources of radiation and probing the reionization era by mapping the 21-cm emission from neutral regions. While its energy density is estimated to be only a 1% correction to that of the CMB, the redshifted 21-cm emission should display angular structure as well as frequency structure due to inhomogeneities in the gas density field, hydrogen ionized fraction, and spin temperature. Indeed, a full mapping of the distribution of H I as a function of redshift is possible in principle.

The basic physics of the hydrogen spin transition is determined as follows. The ground-state hyperfine levels of hydrogen tend to thermalize with the CMB background, making the IGM unobservable. If other processes shift the hyperfine level populations away from thermal equilibrium, then the gas becomes observable against the CMB in emission or in absorption. The relative occupancy of the spin levels is usually described in terms of the hydrogen spin temperature T_S, defined by

$$\frac{n_1}{n_0} = 3\,\exp\left\{-\frac{T_*}{T_S}\right\} \ , \tag{40}$$

where n_0 and n_1 refer respectively to the singlet and triplet hyperfine levels in the atomic ground state ($n = 1$), and $T_* = 0.068$ K is defined by $k_B T_* = E_{21}$, where the energy of the 21 cm transition is $E_{21} = 5.9 \times 10^{-6}$ eV, corresponding to a frequency of 1420 MHz. In the presence of the CMB alone, the spin states reach thermal equilibrium with $T_S = T_{\rm CMB} = 2.725(1+z)$ K on a time-scale of $T_*/(T_{\rm CMB} A_{10}) \simeq 3 \times 10^5 (1+z)^{-1}$ yr, where $A_{10} = 2.87 \times 10^{-15}$ s^{-1} is the spontaneous decay rate of the hyperfine transition. This time-scale is much shorter than the age of the universe at all redshifts after cosmological recombination.

The IGM is observable when the kinetic temperature T_k of the gas differs from $T_{\rm CMB}$ and an effective mechanism couples T_S to T_k. Collisional de-excitation of the triplet level dominates at very high redshift, when the gas density (and thus the collision rate) is still high, but once a significant galaxy population forms in the universe, the spin temperature is affected also by an indirect mechanism that acts through the scattering of Lyman-α photons. Continuum UV photons produced by early radiation sources redshift by the Hubble expansion into the local Lyman-α line at a lower redshift. These photons mix the spin states via the Wouthuysen-Field process whereby an atom initially in the $n = 1$ state absorbs a Lyman-α photon, and the spontaneous decay which returns it from $n = 2$ to $n = 1$ can result in a final spin state which is different from the initial one. Since the neutral IGM is highly opaque to resonant scattering, and the Lyman-α photons receive Doppler kicks in each scattering, the shape of the radiation spectrum near Lyman-α is determined by T_k, and the resulting spin temperature (assuming $T_S \gg T_*$) is then a weighted average of T_k and $T_{\rm CMB}$:

$$T_S = \frac{T_{\rm CMB} T_k (1 + x_{\rm tot})}{T_k + T_{\rm CMB} x_{\rm tot}} \ , \tag{41}$$

where $x_{\rm tot} = x_\alpha + x_c$ is the sum of the radiative and collisional threshold parameters. These parameters are

$$x_\alpha = \frac{P_{10} T_\star}{A_{10} T_{\rm CMB}} \ , \tag{42}$$

and

$$x_c = \frac{4\kappa_{1-0}(T_k)\, n_H T_\star}{3 A_{10} T_{\rm CMB}} \ , \tag{43}$$

where P_{10} is the indirect de-excitation rate of the triplet $n = 1$ state via the Wouthuysen-Field process, related to the total scattering rate P_α of Lyman-α photons by $P_{10} = 4P_\alpha/27$. Also, the atomic coefficient $\kappa_{1-0}(T_k)$ is tabulated as a function of T_k. The coupling of the spin temperature to the gas temperature becomes substantial when $x_{\rm tot} > 1$; in particular, $x_\alpha = 1$ defines the thermalization rate of P_α:

$$P_{\rm th} \equiv \frac{27 A_{10} T_{\rm CMB}}{4 T_*} \simeq 7.6 \times 10^{-12} \left(\frac{1+z}{10}\right) \ {\rm s}^{-1} \ . \tag{44}$$

A patch of neutral hydrogen at the mean density and with a uniform T_S produces (after correcting for stimulated emission) an optical depth at a present-day (observed) wavelength of $21(1+z)$ cm,

$$\tau(z) = 9.0 \times 10^{-3} \left(\frac{T_{\rm CMB}}{T_S}\right) \left(\frac{\Omega_b h}{0.03}\right) \left(\frac{\Omega_m}{0.3}\right)^{-1/2} \left(\frac{1+z}{10}\right)^{1/2} \ , \tag{45}$$

assuming a high redshift $z \gg 1$. The observed spectral intensity I_ν relative to the CMB at a frequency ν is measured by radio astronomers as an effective brightness temperature T_b of blackbody emission at this frequency, defined using the Rayleigh-Jeans limit of the Planck radiation formula: $I_\nu \equiv 2 k_B T_b \nu^2 / c^2$.

The brightness temperature through the IGM is $T_b = T_{\rm CMB} e^{-\tau} + T_S (1 - e^{-\tau})$, so the observed differential antenna temperature of this region relative to the CMB is

$$\begin{aligned} T_b &= (1+z)^{-1} (T_S - T_{\rm CMB})(1 - e^{-\tau}) \\ &\simeq 28\,{\rm mK} \left(\frac{\Omega_b h}{0.033}\right) \left(\frac{\Omega_m}{0.27}\right)^{-1/2} \left(\frac{1+z}{10}\right)^{1/2} \left(\frac{T_S - T_{\rm CMB}}{T_S}\right) \ , \end{aligned} \tag{46}$$

where $\tau \ll 1$ is assumed and T_b has been redshifted to redshift zero. Note that the combination that appears in T_b is

$$\frac{T_S - T_{\rm CMB}}{T_S} = \frac{x_{\rm tot}}{1 + x_{\rm tot}} \left(1 - \frac{T_{\rm CMB}}{T_k}\right) \ . \tag{47}$$

In overdense regions, the observed T_b is proportional to the overdensity, and in partially ionized regions T_b is proportional to the neutral fraction. Also, if $T_S \gg T_{\rm CMB}$ then the IGM is observed in emission at a level that is independent of T_S. On the other hand, if $T_S \ll T_{\rm CMB}$ then the IGM is observed in absorption at a level that is enhanced by a factor of $T_{\rm CMB}/T_S$. As a result, a number of cosmic events are expected to leave observable signatures in the redshifted 21-cm line, as discussed below in further detail.

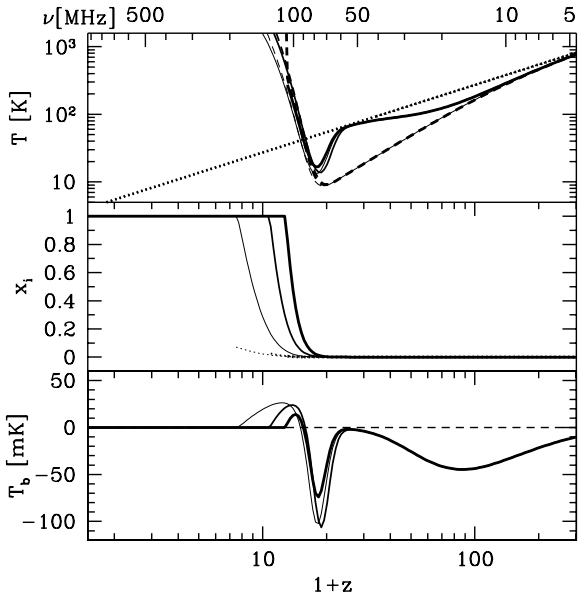

Figure 10: *Top panel:* Evolution with redshift z of the CMB temperature $T_{\rm CMB}$ (dotted curve),the gas kinetic temperature T_k (dashed curve), and the spin temperature T_S (solid curve). *Middle panel:* Evolution of the gas fraction in ionized regions x_i (solid curve) and the ionized fraction outside these regions (due to diffuse X-rays) x_e (dotted curve). *Bottom panel:* Evolution of mean 21 cm brightness temperature T_b. The horizontal axis at the top provides the observed photon frequency at the different redshifts shown at the bottom. Each panel shows curves for three models in which reionization is completed at different redshifts, namely $z = 6.47$ (thin curves), $z = 9.76$ (medium curves), and $z = 11.76$ (thick curves). **Figure credit:** Pritchard, J., & Loeb, A., *Phys. Rev.* **D78**, 3511 (2008).

Figure 10 illustrates the mean IGM evolution for three examples in which reionization is completed at different redshifts, namely $z = 6.47$ (thin curves), $z = 9.76$ (medium curves), and $z = 11.76$ (thick curves). The top panel shows the global evolution of the CMB temperature $T_{\rm CMB}$ (dotted curve), the gas kinetic temperature T_k (dashed curve), and the spin temperature T_S (solid curve). The middle panel shows the evolution of the ionized gas fraction and the bottom panel presents the mean 21 cm brightness temperature, T_b.

3.2.2 A handy tool for studying cosmic reionization

The prospect of studying reionization by mapping the distribution of atomic hydrogen across the universe using its prominent 21-cm spectral line has motivated several teams to design and construct arrays of low-frequency radio telescopes; the Low Frequency Array (http://www.lofar.org/), the Murchison Wide-Field Array (*http://www.mwatelescope.org/*), PAPER (*http://arxiv.org/abs/0904.1181*), GMRT (*http://arxiv.org/abs/0807.1056*), 21CMA (*http://21cma.bao.ac.cn/*), and ultimately the Square Kilometer Array (*http://www.skatelescope.org*) will search over the next decade for 21-cm emission or absorption from $z \sim 6.5$–15, redshifted and observed today at relatively low frequencies which correspond to wavelengths of 1.5 to 4 meters.

The idea is to use the resonance associated with the hyperfine splitting in the ground state of hydrogen. While the CMB spectrum peaks at a wavelength of 2 mm, it provides a still-measurable intensity at meter wavelengths that can be used as the bright background source against which we can see the expected 1% absorption by neutral hydrogen along the line of sight. The hydrogen gas produces 21-cm absorption if its spin temperature is colder than the CMB and excess emission if it is hotter. Since the CMB covers the entire sky, a complete three-dimensional map of neutral hydrogen can in principle be made from the sky position of each absorbing gas cloud together with its redshift z. Different observed wavelengths slice the Universe at different redshifts, and ionized regions are expected to appear as cavities in the hydrogen distribution, similar to holes in swiss cheese. Because the smallest angular size resolvable by a telescope is proportional to the observed wavelength, radio astronomy at wavelengths as large as a meter has remained relatively undeveloped. Producing resolved images even of large sources such as cosmological ionized bubbles requires telescopes which have a kilometer scale. It is much more cost-effective to use a large array of thousands of simple antennas distributed over several kilometers, and to use computers to cross-correlate the measurements of the individual antennas and combine them effectively into a single large telescope. The new experiments

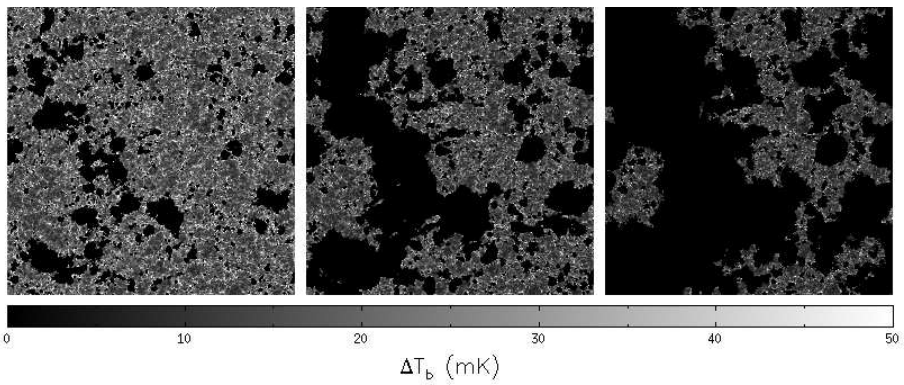

Figure 11: Map of the fluctuations in the 21 cm brightness temperature on the sky, ΔT_b (mK), based on a numerical simulation which follows the dynamics of dark matter and gas in the IGM as well as the radiative transfer of ionizing photons from galaxies. The panels show the evolution of the signal in a slice of 140 comoving Mpc on a side, in three snapshots corresponding to the simulated volume being 25, 50, and 75 % ionized. Since neutral regions correspond to strong emission (i.e., a high T_b), the 21-cm maps illustrate the global progress of reionization and the substantial large-scale spatial fluctuations in the reionization history. **Figure credit:** Trac, H., Cen, R., & Loeb, A., *Astrophys. J.* **689**, L81 (2009).

are being placed mostly in remote sites, because the cosmic wavelength region overlaps with more mundane terrestrial telecommunications.

In approaching redshifted 21-cm observations, although the first inkling might be to consider the mean emission signal in the bottom panel of Figure 10, the signal is orders of magnitude fainter than foreground synchrotron emission from relativistic electrons in the magnetic field of our own Milky Way as well as other galaxies (see Figure 12). Thus cosmologists have focused on the expected characteristic variations in T_b, both with position on the sky and especially with frequency, which signifies redshift for the cosmic signal. The synchrotron foreground is expected to have a smooth frequency spectrum, and so it is possible to isolate the cosmological signal by taking the difference in the sky brightness fluctuations at slightly different frequencies (as long as the frequency separation corresponds to the characteristic size of ionized bubbles). The 21-cm brightness temperature depends on the density of neutral hydrogen. As explained in the previous subsection, large-scale patterns in the reionization are driven by spatial variations in the abundance of galaxies; the 21-cm fluctuations reach $\sim$5 mK (root mean square) in brightness temperature on a scale of 10 comoving Mpc. While detailed maps will be difficult to extract due to the foreground emission, a statistical detection of these fluctuations is expected to be well within the capabilities of the first-generation experiments now being built. Current work suggests that the key information on the topology and timing of reionization can be extracted statistically.

While numerical simulations of reionization are now reaching the cosmological box sizes needed to predict the large-scale topology of the ionized bubbles, they do this at the price of limited small-scale resolution (see Figure 11). These simulations cannot yet follow in any detail the formation of individual stars within galaxies, or the feedback that stars produce on the surrounding gas, such as photo-heating or the hydrodynamic and chemical impact of supernovae, which blow hot bubbles of gas enriched with the chemical products of stellar nucleosynthesis. Thus, the simulations cannot directly predict whether the stars that form during reionization are similar to the stars in the Milky Way and nearby galaxies or to the primordial $100M_\odot$ stars. They also cannot determine whether feedback prevents low-mass dark matter halos from forming stars. Thus, models are needed that make it possible to vary all these astrophysical parameters of the ionizing sources and to study the effect on the 21-cm observations.

The theoretical expectations presented here for reionization and for the 21-cm signal are based on rather large extrapolations from observed galaxies to deduce the properties of much smaller galaxies that formed at an earlier cosmic epoch. Considerable surprises are thus possible, such as an early population of quasars or even unstable exotic particles that emitted ionizing radiation as they decayed. In any case, the forthcoming observational data in 21-cm cosmology should make the next few years a very exciting time.

At high redshifts prior to reionization, spatial perturbations in the thermodynamic gas properties are linear and can be predicted precisely (see section 2.1). Thus, if the gas is probed with the 21-cm technique then it becomes a promising tool of fundamental, precision cosmology, able to probe the primordial power spectrum of density fluctuations imprinted in the very early universe, perhaps in an era of cosmic inflation. The 21-cm fluctuations can be measured down to the smallest scales where the baryon pressure suppresses gas fluctuations, while the CMB anisotropies are damped on small scales (through the so-called Silk damping).

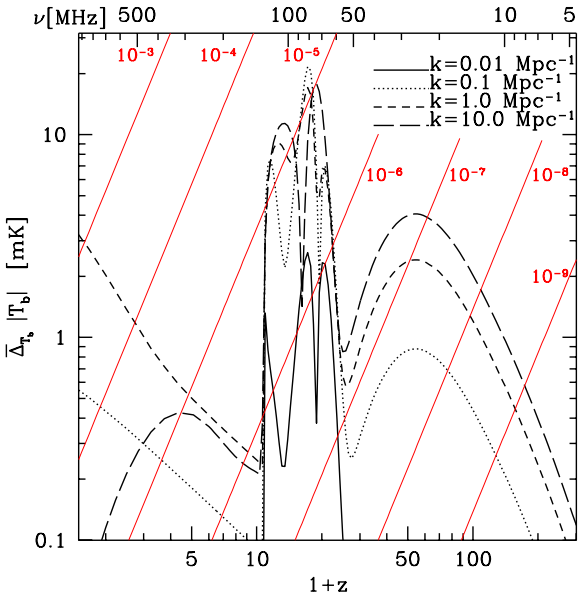

Figure 12: Predicted redshift evolution of the angle-averaged amplitude of the 21-cm power spectrum ($|\bar{\Delta}_{T_b}| = [k^3 P_{21-\rm cm}(k)/2\pi^2]^{1/2}$) at comoving wavenumbers $k = 0.01$ (solid curve), 0.1 (dotted curve), 1.0 (short dashed curve), 10.0 (long dashed curve), and 100.0Mpc^{-1} (dot-dashed curve). In the model shown, reionization is completed at $z = 9.76$. The horizontal axis at the top shows the observed photon frequency at the different redshifts. The diagonal straight (red) lines show various factors of suppression for the synchrotron Galactic foreground, necessary to reveal the 21-cm signal. **Figure credit:** Pritchard, J.R. & Loeb A., *Phys. Rev* **D78**, 3511 (2008).

This difference in damping scales can be seen by comparing the baryon-density and photon-temperature power spectra. Since the 21-cm technique is also three-dimensional (while the CMB yields a single sky map), there is a much large potential number of independent modes probed by the 21-cm signal: $N_{21-\rm cm} \sim 3 \times 10^{16}$ compared to $N_{\rm cmb} \sim 2 \times 10^7$. This larger number should provide a measure of non-Gaussian deviations to a level of $\sim N_{21\rm cm}^{-1/2}$, constituting a test of the inflationary origin of the primordial inhomogeneities which are expected to possess non-Gaussian deviations $> 10^{-6}$.

The 21cm fluctuations are expected to simply trace the primordial power-spectrum of matter density perturbations (which is shaped by the initial conditions from inflation and the dark matter) either before the first population of galaxies had formed (at redshifts $z > 25$) or after reionization ($z < 6$) – when only dense pockets of self-shielded hydrogen (such as damped Lyman-α systems) survive. During the epoch of reionization, the fluctuations are mainly shaped by the topology of ionized regions, and thus depend on uncertain astrophysical details involving star formation. However, even during this epoch, the imprint of peculiar velocities (which are induced gravitationally by density fluctuations), can in principle be used to separate the implications for fundamental physics from the astrophysics.

Peculiar velocities imprint a particular form of anisotropy in the 21-cm fluctuations that is caused by gas motions along the line of sight. This anisotropy, expected in any measurement of density that is based on a spectral resonance or on redshift measurements, results from velocity compression. Consider a photon traveling along the line of sight that resonates with absorbing atoms at a particular point. In a uniform, expanding universe, the absorption optical depth encountered by this photon probes only a narrow strip of atoms, since the expansion of the universe makes all other atoms move with a relative velocity that takes them outside the narrow frequency width of the resonance line. If there is a density peak, however, near the resonating position, the increased gravity will reduce the expansion velocities around this point and bring more gas into the resonating velocity width. This effect is sensitive only to the line-of-sight component of the velocity gradient of the gas, and thus causes an observed anisotropy in the power spectrum even when all physical causes of the fluctuations are statistically isotropic. This anisotropy is particularly important in the case of 21-cm fluctuations. When all fluctuations are linear, the 21-cm power spectrum takes the form

$$P_{21-\rm cm}(\mathbf{k}) = \mu^4 P_\rho(k) + 2\mu^2 P_{\rho-\rm iso}(k) + P_{\rm iso} \ , \tag{48}$$

where $\mu = \cos\theta$ in terms of the angle θ between the wave-vector $\mathbf{k}$ of a given Fourier mode and the line of sight, $P_{\rm iso}$ is the isotropic power spectrum that would result from all sources of 21-cm fluctuations without velocity compression, $P_\rho(k)$ is the 21-cm power spectrum from gas density fluctuations alone, and $P_{\rho-\rm iso}(k)$

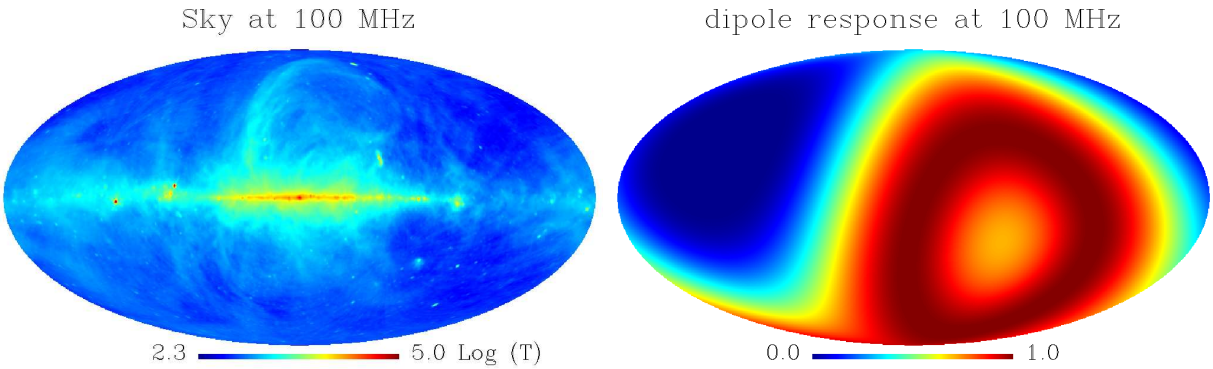

Figure 13: *Left panel:* Radio map of the sky at 100 MHz. *Right panel:* Ideal dipole response averaged over 24 hours. **Figure credits:** Pritchard, J., & Loeb, A. Phys. Rev. **D**, in press (2010); de Oliveira-Costa, A. et al. Mon. Not. R. Astron. Soc. **388**, 247 (2008).

is the Fourier transform of the cross-correlation between the density and all sources of 21-cm fluctuations. The three power spectra can also be denoted $P_{\mu^4}(k)$, $P_{\mu^2}(k)$, and $P_{\mu^0}(k)$, according to the power of μ that multiplies each term. At these redshifts, the 21-cm fluctuations probe the infall of the baryons into the dark matter potential wells. The power spectrum shows remnants of the photon-baryon acoustic oscillations on large scales, and of the baryon pressure suppression on small scales.

Once stellar radiation becomes significant, many processes can contribute to the 21-cm fluctuations. The contributions include fluctuations in gas density, temperature, ionized fraction, and Lyα flux. These processes can be divided into two broad categories: The first, related to *"physics"*, consists of probes of fundamental, precision cosmology, and the second, related to *"astrophysics"*, consists of probes of stars. Both categories are interesting – the first for precision measures of cosmological parameters and studies of processes in the early universe, and the second for studies of the properties of the first galaxies. However, the astrophysics depends on complex non-linear processes (collapse of dark matter halos, star formation, supernova feedback), and must be cleanly separated from the physics contribution, in order to allow precision measurements of the latter. As long as all the fluctuations are linear, the anisotropy noted above allows precisely this separation of the *fundamental physics* from the *astrophysics* of the 21-cm fluctuations. In particular, the $P_{\mu^4}(k)$ is independent of the effects of stellar radiation, and is a clean probe of the gas density fluctuations. Once non-linear terms become important, there arises a significant mixing of the different terms; in particular, this occurs on the scale of the ionizing bubbles during reionization.

The 21-cm fluctuations are affected by fluctuations in the Lyman-α flux from stars, a result that yields an indirect method to detect and study the early population of galaxies at $z \sim 20$. The fluctuations are caused by biased inhomogeneities in the density of galaxies, along with Poisson fluctuations in the number of galaxies. Observing the power-spectra of these two sources would probe the number density of the earliest galaxies and the typical mass of their host dark matter halos. Furthermore, the enhanced amplitude of the 21-cm fluctuations from the era of Lyα coupling improves considerably the practical prospects for their detection. Precise predictions account for the detailed properties of all possible cascades of a hydrogen atom after it absorbs a photon. Around the same time, X-rays may also start to heat the cosmic gas, producing strong 21-cm fluctuations due to fluctuations in the X-ray flux.

In difference from interferometric arrays, single dipole experiments which integrate over most of the sky, can search for the global (spectral) 21-cm signal shown in Figure 10. Examples of such experiments are CoRE or EDGES (*http://www.haystack.mit.edu/ast/arrays/Edges/*). Rapid reionization histories which span a redshift range $\Delta z < 2$ can be constrained, provided that local foregrounds (see Figure 13) can be well modelled by low-order polynomials in frequency. Observations in the frequency range 50-100 MHz can potentially constrain the Lyman-α and X-ray emissivity of the first stars forming at redshifts $z \sim 15$–25, as illustrated in Figure 14.

4 Epilogue

The initial conditions of our Universe can be summarized on a single sheet of paper. Yet the Universe is full of complex structures today, such as stars, galaxies and groups of galaxies. This chapter discussed the standard theoretical model for how complexity emerged from the simple initial state of the Universe through the action of gravity. In order to test and inform the related theoretical calculations, large-aperture telescopes and arrays of radio antennae are currently being designed and constructed.

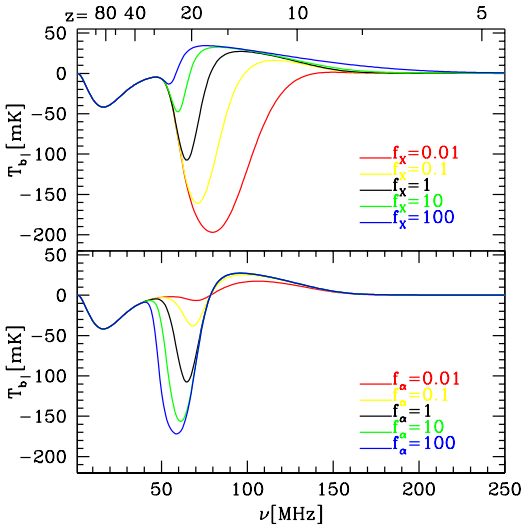

Figure 14: Dependence of global 21-cm signal on the X-ray (top panel) and Lyman-α (bottom panel) emissivity of stars. Each case depicts examples with the characteristic emissivity reduced or increased by a factor of up to 100. **Figure credit:** Pritchard, J., & Loeb, A. Phys. Rev. **D**, in press (2010).

The actual transition from simplicity to complexity has not been observed as of yet. The simple initial conditions were already traced in maps of the microwave background radiation, but the challenge of detecting the first generation of galaxies defines one of the exciting frontiers in the future of cosmology. Once at hand, the missing images of the infant Universe might potentially surprise us and revise our current ideas.

Acknowledgements

I thank my collaborators on the topics covered by this chapter: Dan Babich, Rennan Barkana, Volker Bromm, Steve Furlanetto, Zoltan Haiman, Joey Munoz, Jonathan Pritchard, Hy Trac, Stuart Wyithe, and Matias Zaldarriaga.

Further Reading: Loeb, A., *"How Did the First Stars and Galaxies Form?"*, Princeton University Press (2010).